accenture

Driving intelligence at the edge with Neuromorphic Computing

Brain-inspired computing will expand AI beyond the cloud with adaptable and efficient devices.

Accenture Labs

As artificial intelligence has grown increasingly sophisticated, it's become a crucial element of many products and services. Today's AI systems can interpret spoken commands, recognize objects and gestures, navigate, plan, and make decisions. These successes are driving demand for even more powerful AI-driven experiences: smart products and environments that are autonomous, interact naturally with users, and adapt to changing conditions seamlessly. But keeping up with these increasing expectations will require new thinking about how AI systems are designed.

One promising approach is to turn to the brain for inspiration. The human brain packs an amazing capacity for learning and computation into a compact package. It consumes a fraction of the energy required by the processors that power today's AI systems, and needs only a few examples to learn new patterns.

It's responsive, having evolved to quickly identify and avoid lunging predators. It has many properties we need in the next wave of smart products that provide efficient, responsive, and adaptive intelligence at the edge.

This is where neuromorphic computing comes in. Neuromorphic computing is an emerging approach to hardware-accelerated AI. Inspired by the properties of biological brains, neuromorphic architectures are radically different from those used in traditional processors. Instead, they emulate neural systems.

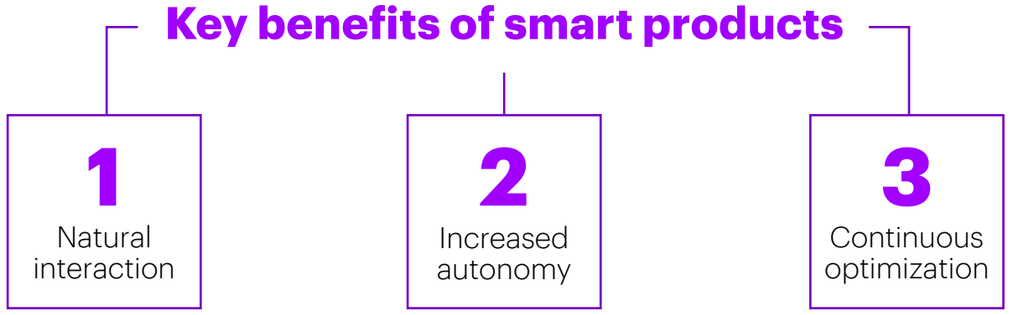
Neuromorphic technologies will help solve business challenges that require AI at the edge, such as responsive voice control for vehicles, full-body gesture recognition for touchless interfaces, and on-board intelligence for assistive robotics.

Powering increasingly sophisticated smart products

Smart products are already disrupting and transforming industries, with growing demand from both consumers and businesses. The smart home market alone, just one piece of the smart product spectrum, is projected to be worth $135B by 2035.[1]

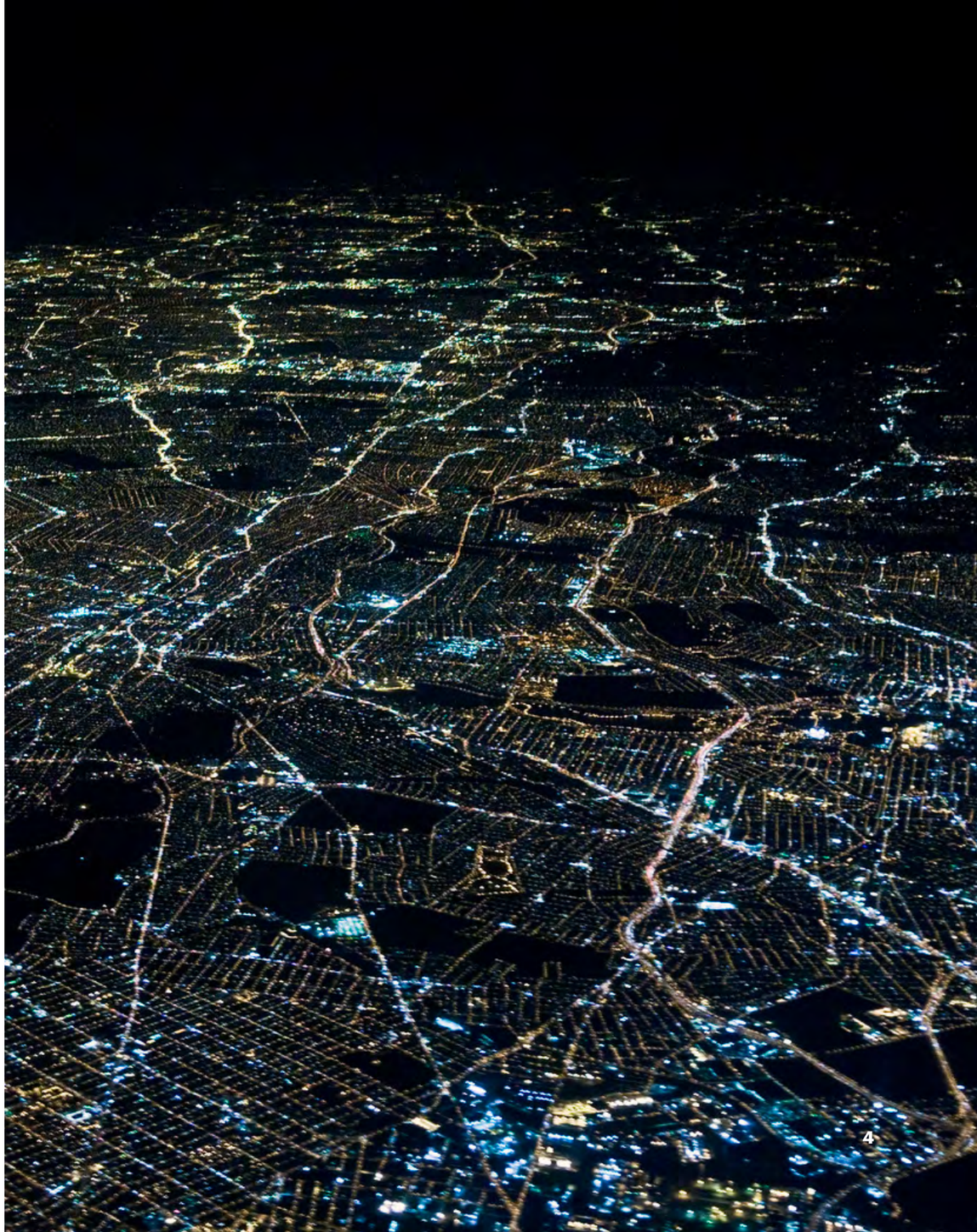
There are several reasons for this trend toward "smart everything." Smart products offer **natural interaction**: many products, from home entertainment to automotive interiors to industrial equipment, can be controlled by voice and gesture rather than physical buttons or control panels. Today's smart products let people focus on other activities by operating with **increased autonomy**, like robot vacuums that clean the house or the aisles of the grocery store.

The most advanced smart products also offer **continuous optimization**. Through enhanced data collection and analysis, they can improve their performance over time to better serve the needs of a particular customer or organization. Whether it's a smart thermostat that learns optimal room temperatures for different situations or a robot that develops a new, more efficient path through a warehouse, optimization is a powerful capability. Optimization also offers a pathway to provide new services via existing smart products. As companies learn how their products are being used, they can identify opportunities for expanded services and features, and roll those services out to expand the capabilities of the product.

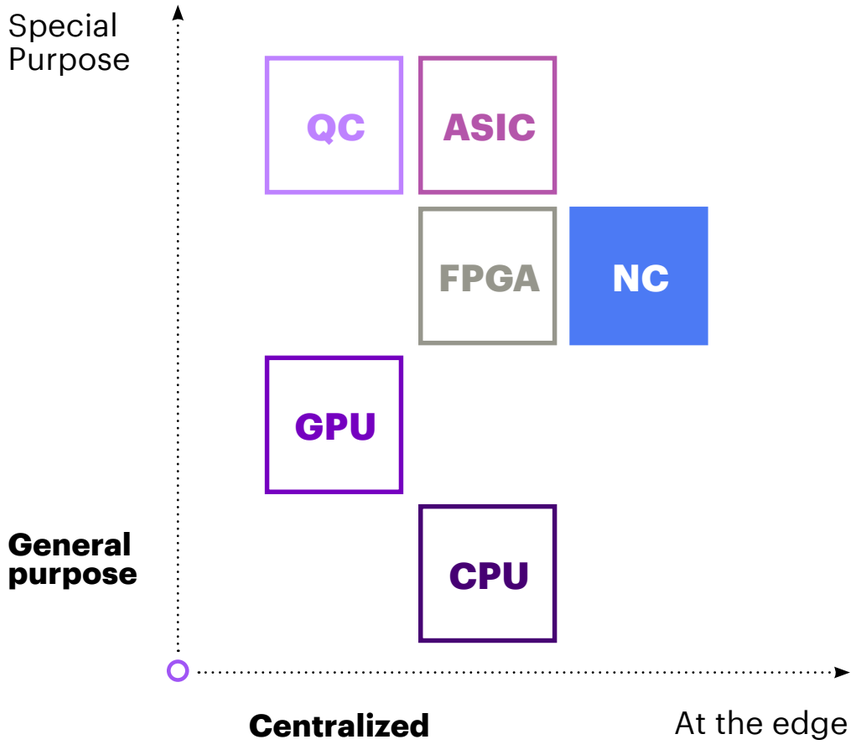


Today’s smart products have demonstrated the value of infusing AI into both consumer and industrial devices. But we’re running up against the limits of traditional computing: the idea of delivering AI everywhere is limited by the hardware that powers it. Tasks like recognizing real-world gestures or understanding speech require powerful processors that quickly drain a device’s batteries.

That’s why many of the smart products we use today actually rely on off-site computing power to run their AI. But this approach means that a product needs a reliable network connection to act smart; it also means introducing delays while the system sends data off-site and waits to find out how it should respond. For some applications this isn’t an issue, and the best modern networks may reduce the latency going forward; but for applications that demand truly real-time performance, any delay may be a dealbreaker. Sending raw data outside the device creates privacy and cybersecurity concerns as well.



Computational variety landscape



As companies look to a wider range of smart applications, these challenges highlight the need for more intelligent computing on-board smart products and devices, at the "edge" of networks. Achieving the full range of capabilities that both businesses and consumers want will require the power of cloud capabilities coupled with **intelligence at the edge**.

To enable "intelligence everywhere," businesses are looking to new computing paradigms. The next era of computing will rely on what Accenture has described as **computational variety**.[2] Computational variety means matching the needs of business applications to specialized computing hardware from a growing set of options. Neuromorphic technologies will be among those options, along with quantum computers, field-programmable gate arrays (FPGAs), and application-specific integrated circuits (ASICs). All of these offer dramatic performance benefits over CPUs and GPUs for particular applications. When the needs of the application include low-power, low-latency, or on-device adaptation, that's where neuromorphic solutions may be the best match.

CASE STUDIES

Applied today:

Responsive voice control for smart vehicles

Owners of smart vehicles have high expectations: they want functions like self-parking and summoning features, but they also want natural, seamless interfaces to control their interactions with the car. Voice-based controls, along with implicit intent recognition, can provide greater personalization and smoother interactions. They also help unclutter the car's dashboards and displays.

Why don't today's vehicles make wider use of this approach? Because it requires intensive computation. Conventional AI hardware is too power-hungry to run onboard continuously without draining the battery, and using only cloud-based AI and wake words creates too much lag, leading to a poor experience.

Neuromorphic technologies make efficient onboard AI possible. In a recent collaboration with an automotive client, we demonstrated that spiking neural networks running on a neuromorphic processor can recognize simple voice commands up to 0.2 seconds faster than a commonly used embedded GPU accelerator, while using up to a thousand times less power. This brings truly intelligent, low latency interactions into play, at the edge, even within the power-limited constraints of a parked vehicle.

AI powered by brain-like computing architectures

Scientific understanding of how the brain works is not yet complete, but it is mature enough to uncover many core principles of neural computation. Researchers and engineers have worked together to develop algorithms and processors that replicate some of those core principles and mechanisms.

What are they trying to emulate? An average human brain contains 80 to 100 billion neurons that are each highly efficient. Activity in the whole brain is much sparser than traditional computer architectures. Complex sequences of spikes in organic nerve fibers are nothing like the 64-bit silicon data buses we see in general-purpose processors. In the brain, each neuron works asynchronously to provide massive parallelism—many different processes all happen at once—and to adapt quickly to rapid changes in the environment.

In the past several years, neuromorphic devices with these properties have become a reality, accelerating practical solutions to the increasing demand for smart products.

These properties of biological brains are at the core of brain-like computing, which provides:

Energy efficiency
Neuromorphic systems are several orders of magnitude more energy efficient than general purpose computing architectures.

Low latency
Neuromorphic systems excel at processing continuous streams of data and deploying neuromorphic processors at the edge reduces the delay to analysis.

Adaptive processing
Neuromorphic system architectures let devices adapt to changes in context.

Rapid learning
Recent advances in training neuromorphic systems have enabled rapid learning from very little data—near-biological capabilities which are beyond most conventional AI systems.

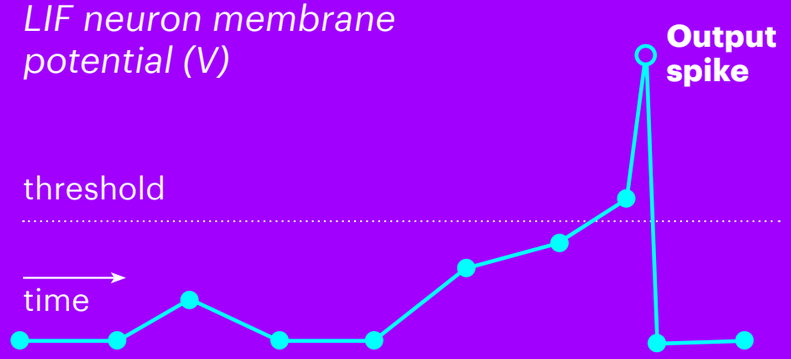
NUTS AND BOLTS

Neuromorphic Computing 101

The term Neuromorphic Computing is used to describe a variety of computational technologies inspired by the brain. We're focused on systems using spiking neural networks, which we see as the most application-ready approaches in neuromorphic computing today.

Spiking neural networks are built from two components: the neuron model and the synapse model. The neuron model emulates the spiking activity of a biological neuron using a digital circuit. That means the neuron either fires a spike, or it's silent. In most computing systems, a large portion of energy is used moving information between processing cores or in and out of memory. But in a spiking neuron model, when a neuron is not spiking, it does not move any information and uses very little energy. SNNs can include millions of neurons, each working independently to aggregate input values and fire output spikes.

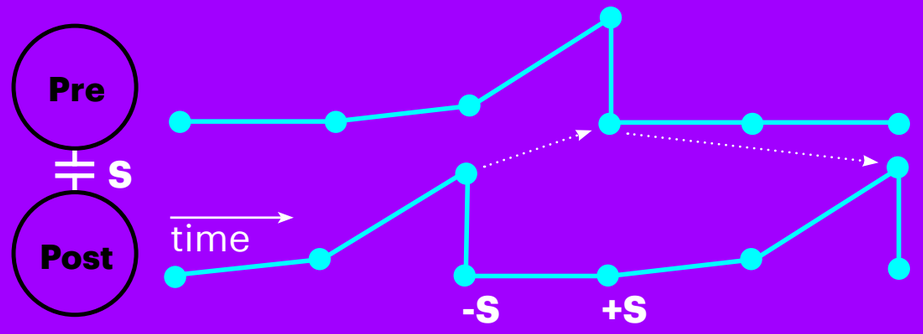
How do spiking neurons work?



A neuron receives input signals from sensors and other neurons over time. The neuron adds the inputs to its membrane potential (V) and subtracts a fraction of the value from the previous timestep (leak). If the integrated input values are large enough, V exceeds the threshold, causing the neuron to fire an output spike and reset V back to zero.

How do spiking networks learn?

Spike timing dependent plasticity (STDP)



Spike-timing dependent plasticity (STDP) enables SNNs to learn. Under STDP, the strength of each synapse is adjusted based on the precise time each connected neuron fires a spike. STDP works by learning associations in the input data. When the sending neuron spikes before the receiving neuron, the synapse is strengthened. If the two neurons spike in reverse order, the synapse is weakened. Variations on STDP can support many types of machine learning, including supervised learning, reinforcement learning, and other AI algorithms.

This approach is very different from the artificial neural networks (ANNs) powering most AI systems today. In those systems, nodes map input values to outputs with simple activation functions. ANNs don't process changing input signals continuously. Instead, input values are sent to the network in batches and the activation of the entire network is re-computed for each set of inputs in a batch. That can require trillions of operations per second, which is why current AI models run on power-hungry GPUs.

In spiking neural networks, connecting neurons is the job of the synapse model. The synapse model takes output spikes from each neuron and routes them to the inputs of other neurons. The synapse scales the signal by a factor called synaptic strength. The SNN learns from experience and adapts to an environment by adjusting synaptic strengths. Again, this is very different from ANNs which use a method called backpropagation to optimize connection strengths for a given problem.

CASE STUDIES

Applied tomorrow:

Adaptive control for semi-autonomous robots

Many promising robotics applications need precise motor control before they're ready for practical use. Precision agriculture drones need to be able to target specific plants. Medical telepresence robots must be capable of navigating safely around patients and other healthcare workers.

Current approaches work well for highly repetitive movements and controlled circumstances. But designing a robot with precise motor control that can also adapt to work in a variety of situations and contexts has so far proven out of reach.

Recent developments show promise in this area. Adaptive control algorithms inspired by the motor control structures of biological brains have shown an impressive ability to move precisely and compensate for a variety of unexpected conditions. These neuromorphic algorithms can also rapidly adapt to new applications. We're working with researchers from the Open University of Israel and ALYN Hospital to apply these algorithms to wheelchair-mounted assistive robot arms. These robots must be extremely precise and also adaptable to a range of daily tasks such as feeding and opening doors. The high cost of existing systems is a significant barrier; we expect the brain-inspired neuromorphic solution to allow a much lower cost robot to support the same range of tasks.

Recent advances put neuromorphic computing within reach

Until recently, spiking neural networks were used in neuroscience but little else.

They can be simulated using CPUs, but this doesn't achieve the low power use and low latency benefits of true neuromorphic computing. Realizing the true potential of neuromorphic computing requires new computing architectures.

In the past decade, research and development on neuromorphic processors, sensors, and algorithms has rapidly advanced with support from the DARPA SyNAPSE project and the EU Human Brain Project. IBM developed the TrueNorth processor and model architecture in 2014, demonstrating that neuromorphic chips could reach a massive scale (more than 1 billion neurons) and that they could efficiently perform state-of-the-art deep learning tasks[3]. A consortium of researchers in the EU developed two neuromorphic platforms starting in 2013; these helped establish that neuromorphic computing could be made practical with inexpensive ARM cores and could push the boundaries of electrical engineering with massive silicon wafer analog circuits.

Intel announced its own neuromorphic architecture in 2017, called Loihi. The new chip has been made available to a growing community of researchers, including Accenture Labs. Current Loihi devices are still intended for research rather than immediate commercialization, but academic and industrial scientists around the world are already applying them to many real-world problems. Benchmarks have shown that the chip uses 40x less energy than a standard GPU-based approach for real-time speech processing, for example.[4]

Sensor startups Prophesee and Inivation have also made significant advances in developing neuromorphic architectures specifically for computer vision. Event-based cameras developed by sensor startups Prophesee and Inivation are both massively parallel, asynchronous, spiking sensors that provide drastically lower energy consumption, lower latency, and higher dynamic range than standard image sensing chips.

CASE STUDIES

Embracing change:

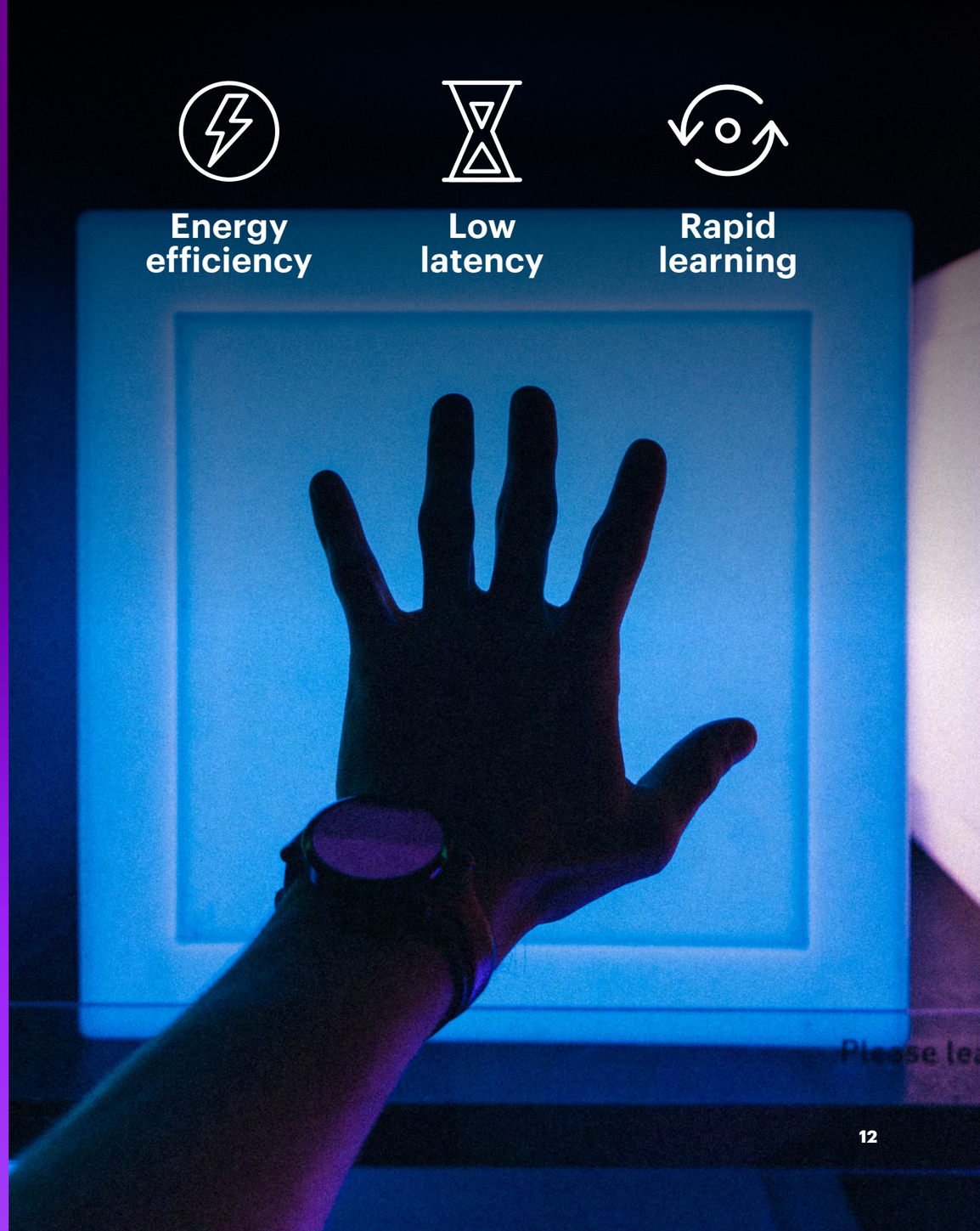
Flexible gesture recognition for touchless interaction

It's now common to interact with systems through touch interfaces, like retail payment touchscreens and interactive displays. But high-touch surfaces like these have their limitations, and when they're at high use in public spaces, they can also spread germs. Enabling more flexible and touchless gesture-based interaction can protect health as well as creating richer, more natural customer experiences.

Shoppers could interact with smart retail kiosks to learn about products with simple gestures. Movie-goers could engage with dynamic movie posters with a wave and a nod. There are many possibilities; but recognizing gestures in the real world is difficult. Natural gestures vary tremendously between people, and even for a given person, gestures can shift quickly during interactions to reduce effort and improve communication. Humans adapt to these differences easily, but current AI hardware can't. Enter neuromorphic computing.

By pairing a neuromorphic processor with a spiking image sensor, Accenture Labs has developed AI models that support real-time natural gesture recognition. Unlike current AI solutions that would require large amounts of training data to recognize just a few gestures, these models can learn from new input data in real time. The system can quickly learn multiple different gestures from one person, and it can easily recognize different people's gestures as well.

Supporting this level of natural interaction will not only enable safer interactions with technology, it will expand the possibilities for gesture-driven experiences.



The event-based sensor (right) records only when a scene changes. This generates less data, improving energy efficiency and reducing latency in addition to generating more contrast. In the top images, pedestrians passing in front of the sensor are clear even with minimal ambient light. In the bottom images captured from a moving vehicle, the outlines of all relevant objects are distinctly visible. (Images courtesy of Prophesee)

A maturing technology

We've seen a lot of progress in scaling and industrialization of neuromorphic architectures. Still, building and deploying complete neuromorphic solutions will require overcoming some additional challenges.

First, Spiking Neural Networks require new approaches to machine learning. Existing algorithms for supervised learning aren't directly compatible with the mathematically discontinuous nature of spikes. Several research groups have recently developed algorithms to modify the feedback sent to neurons, helping to address this challenge; in the process, they've achieved state-of-the-art accuracy with SNNs on speech and gesture recognition tasks, which are promising steps forward for broader applicability. Now the engineering community must build on this effort with testing and maintenance of these algorithms to expand their suitability for practical use.

As the hardware and algorithmic foundations of neuromorphic computing are coming together to support practical applications of the technology, the next step will be to enable its adoption at scale. We want the tools for developing, debugging, and deploying neuromorphic solutions to be as robust and user-friendly as the tools and APIs used for traditional AI hardware.

We're beginning to see progress in this space: Intel has built a robust set of tools around the Loihi processor, as well as an active user community. Another success in this area has come from Applied Brain Research, a Canadian startup developing a neuromorphic platform that unifies development and deployment across several hardware accelerators. Nevertheless, established machine learning platforms enjoy a considerable lead in this area—for now.

Looking forward

Neuromorphic computing can provide significant business benefits as demand for AI at the edge continues to grow.

Neuromorphic processors and sensors will fill an important niche in the AI landscape: providing real-time intelligence with continuous, onboard learning on a tight energy budget. Already-established use cases include adaptive robotics and advanced features for smart vehicles, but applications will expand considerably as consumers and businesses become more comfortable with machines that learn on the job.

We can also see neuromorphic computing having an impact in cloud and high-performance computing on a longer time horizon. Already, state-of-the-art AI models in the field of natural language processing require so much data and computing power to train that they are out of reach for all but the largest tech firms, and they use staggering amounts of energy. A recent project that trained a robotic hand to manipulate a Rubik's Cube was estimated to require as much as 2.8 GWh of electricity —enough to power hundreds of homes for a year.[5] While current neuromorphic devices aren't yet able to compete with the scale of these massive systems, the potential to reduce those power budgets will motivate further research and development.

Finally, the growing ecosystem of academic, industry, and government participants in neuromorphic computing is a strong positive signal. At Accenture Labs, we're collaborating with leading universities and enterprise partners to envision, invent, and evaluate applications for neuromorphic technologies.

Every organization needs to shape its computational variety strategy to meet growing demands from consumers—and to stay ahead of increasing competition. Now, with emerging neuromorphic hardware and maturing platforms, it's time to start experimenting with neuromorphic computing, starting with applications that require efficient, responsive, and adaptive AI at the edge.

Contacts:

Alex Kass

Fellow and Principal Director, Future Technologies, Accenture Labs

alex.kass@accenture.com

Timothy Shea

Research Lead, Neuromorphic Computing, Accenture Labs

timothy.m.shea@accenture.com

References

1 https://www.marketsandmarkets.com/PressReleases/global-smart-homes-market.asp

2 https://www.accenture.com/us-en/insights/technology/computational-variety

3 https://www.research.ibm.com/articles/brain-chip.shtml

4 https://dl.acm.org/doi/abs/10.1145/3320288.3320304

5 https://www.wired.com/story/ai-great-things-burn-planet/

About Accenture

Accenture is a global professional services company with leading capabilities in digital, cloud and security. Combining unmatched experience and specialized skills across more than 40 industries, we offer Strategy and Consulting, Interactive, Technology and Operations services—all powered by the world's largest network of Advanced Technology and Intelligent Operations centers. Our 514,000 people deliver on the promise of technology and human ingenuity every day, serving clients in more than 120 countries. We embrace the power of change to create value and shared success for our clients, people, shareholders, partners and communities. Visit us at **www.accenture.com**.

About Accenture Labs

Accenture Labs incubates and prototypes new concepts throughapplied R&D projects that are expected to have a significant impact on business and society. Our dedicated team of technologists and researchers work with leaders across the company and external partners to imagine and invent the future.

Accenture Labs is located in seven key research hubs around the world: San Francisco, CA; Washington, D.C.; Dublin, Ireland; Sophia Antipolis, France; Herzliya, Israel; Bangalore, India; Shenzhen, China and Nano Labs across the globe. The Labs collaborates extensively with Accenture's network of nearly 400 innovation centers, studios and centers of excellence to deliver cutting edge research, insights and solutions to clients where they operate and live. For more information, please visit **www.accenture.com/labs**